

**THE MULTIPLE-SERVER QUEUE WITH
HETEROGENEOUS SERVICE TIMES**

A THESIS

Presented to

**The Faculty of the Division of Graduate
Studies and Research**

By

Robert Van Namee Baxley, Jr.

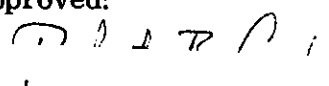
**In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the School
of Industrial and Systems Engineering**

Georgia Institute of Technology

June, 1973

THE MULTIPLE-SERVER QUEUE WITH
HETEROGENEOUS SERVICE TIMES


Approved:



Robert B. Cooper, Chairman



Jamie A. Goode



R. G. Parker

Date approved by Chairman: May 16, 1973

ACKNOWLEDGMENTS

The author wishes to express his sincere thanks to Dr. Robert B. Cooper, who acted as thesis advisor, and who provided the topic which was studied herein. Dr. Cooper has also explored this topic in detail, resulting in his research paper which is listed in the references (7). Although this thesis was completed without reference to that paper, there were several extended conversations with Dr. Cooper which added greatly to the author's understanding of the subject.

Dr. Jamie J. Goode and Dr. R. G. Parker also served on the advisory committee and made valuable suggestions.

Finally, the author's wife, Ginny, deserves special thanks for making this research effort possible.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	ii
LIST OF EXHIBITS.	v
Chapter	
I. INTRODUCTION AND SUMMARY	1
II. LITERATURE SURVEY.	4
Heterogeneous Service Times and Ordered Hunt	
Heterogeneous Service Times and Random	
Choice of Server	
Homogeneous Service Times and Ordered Hunt	
III. STATE EQUATION SOLUTION METHOD FOR BLOCKED CUSTOMERS CLEARED CASE	9
Development of State Space	
Matrix State Equations for Calculating the	
Stationary State Distribution	
Explicit Solution for the Two-Server Problem	
Conclusions	
IV. ALTERNATIVE SOLUTION FOR BLOCKED CUSTOMERS CLEARED CASE	16
Development of Alternate Solution Methods	
Assessment of the Efficiency of these Computations	
Another Method of Calculating the Utilization of	
Each Server	
Alternate Solution for Two-Server Problem	
V. EXTENSION OF RESULTS TO A SYSTEM WITH QUEUEING ALLOWED	29
Definitions and Assumptions	
Construction of Transition Matrix Generator	
Calculation of Results for Generalized Queueing Case	

TABLE OF CONTENTS (Continued)

Chapter	Page
An Interesting Probabilistic Interpretation The Special Case of Blocked Customers Delayed	
VI. RECOMMENDATIONS FOR FURTHER STUDY.	42
APPENDIX.	44
REFERENCES	55

LIST OF EXHIBITS

Exhibit		Page
I.	Calculating $\Psi_m(\mu_m)$ for $s = 4$	27
II.	The Matrix Generator, \hat{G}	33
III.	Diagram of an Arbitrary Cycle of Length X	46

CHAPTER I

INTRODUCTION AND SUMMARY OF RESULTS

The objective of this thesis will be to calculate (1) the utilization of each server and (2) the probability that an arriving customer finds all servers busy for an s server queueing system with heterogeneous service times and ordered hunt. In this context, heterogeneous service times and ordered hunt are defined as follows:

Heterogeneous service times--All service times are independent, negative exponentially distributed random variables, but the mean service time can be different for each server.

Ordered hunt--The servers are numbered and arriving customers who find more than one idle server occupy that idle server with the lowest number.

All calculations are made under the assumption that the system is in a state of statistical equilibrium.

Initially, it is assumed that customers who find all servers busy leave the system, an assumption which will be called "blocked customers cleared" (BCC), and that arrivals follow a Poisson distribution. Traditional methods of writing and solving equations for the stationary state probabilities are examined, and it is concluded that these systems of equations are large and have no simple closed-form solution.

An alternative approach then relaxes the assumption of Poisson arrivals and

considers each server as a one server queueing system with blocked customers cleared and an arrival stream comprised of those customers who overflow all lower-numbered servers. It is demonstrated that the Laplace-Stieltjes transform of the $m + 1^{\text{st}}$ server's interarrival time distribution function, $\Psi_{m+1}(z)$, is related as follows to the corresponding transform for the m^{th} server:

$$\Psi_{m+1}(z) = \frac{\Psi_m(z + \mu_m)}{1 - \Psi_m(z) + \Psi_m(z + \mu_m)} ,$$

where μ_m is the service completion rate for the m^{th} server. Also, the probability that a customer overflowing the first m servers finds the $m+1^{\text{st}}$ server busy is shown to equal $\Psi_{m+1}(\mu_{m+1})$, which can be numerically evaluated by recurrence. Finally, it is demonstrated that the following equations can be used to calculate p_m , the utilization of the m^{th} server, and $A(s)$, the probability that an arriving customer finds all s servers busy:

$$p_m = \begin{cases} \frac{\lambda}{\mu_m} [1 - \Psi_m(\mu_m)] , & m = 1 \\ \frac{\lambda}{\mu_m} \left[\prod_{i=1}^{m-1} \Psi_i(\mu_i) \right] [1 - \Psi_m(\mu_m)] , & m = 2, 3, \dots, s \end{cases}$$

$$A(s) = \prod_{i=1}^s \Psi_i(\mu_i) ,$$

where λ is defined as the mean arrival rate to the first server.

In order to analyze the general case with queueing allowed, it initially seems that the above results no longer apply since a customer overflowing the m^{th} server

can have a later effect on that server. However, if the arrivals are once again assumed to be Poisson, it is demonstrated that the utilization of the m^{th} server and $C(s)$, the probability that all servers are busy, can be related to the corresponding quantities for the BCC case as follows:

$$\hat{p}_m = cp_m + (1 - c) \quad , \quad m = 1, 2, \dots, s$$

$$C(s) = cB(s) + (1 - c) \quad ,$$

where $B(s)$ is the probability that at any instant all servers are busy in a BCC system. It is noted that for Poisson input, the knowledge of whether or not the system is at an arrival instant has no effect on any of the state probabilities so that

$$B(s) = A(s) \quad .$$

The constant c is shown to be equal to the probability that there are no customers waiting in the queue, and it is expressed as follows:

$$c = \left[1 + B(s) \sum_{k=1}^{\infty} \frac{\lambda}{\beta_1 \beta_2 \dots \beta_k} \right]^{-1} \quad ,$$

where β_k is the downward transition rate when all servers are busy and k customers are waiting.

CHAPTER II

LITERATURE SURVEY

With the assumptions of heterogeneous service times and ordered hunt, this multiserver queueing problem has received very little attention in the literature. However, some closely related problems have been analyzed much more thoroughly, and these can be classified as follows:

1. Heterogeneous service times with customers choosing among idle servers at random.
2. Homogeneous service times with ordered hunt.

The following discussion will consider first the work previously done on the problem of heterogeneous service times and ordered hunt, and then it will cover some of the literature on the two related problems.

Heterogeneous Service Times and Ordered Hunt

Only one previous attempt to solve this problem has been discovered. This work, by Vijendra P. Singh, has resulted in two papers. In the first (17) Singh defines a state space for a two-server, heterogeneous system with Poisson arrivals and queueing allowed and then calculates all state probabilities and $E(Q)$, the expected number of customers in the system. He then fixes the mean arrival rate, λ , and the sum of the mean service completion rates, $\mu_1 + \mu_2$, and determines the exact values of μ_1 and μ_2 which will minimize $E(Q)$. After finding these optimum values, he concludes that $E(Q)$ is smaller for the resulting heterogeneous system

than for the corresponding homogeneous whose servers each have the following mean rate, μ :

$$\mu = \frac{\mu_1 + \mu_2}{2}$$

The second paper (18) extends this analysis to the case of 3 servers, but the increased size of the state space makes algebraic solutions much more cumbersome, and therefore numerical solution methods are used. The solution methods used by Singh in these two papers will be explored in greater detail in Chapter III.

Heterogeneous Service Times and Random Choice of Server

In this context, random choice of server means that the arriving customer occupies each idle server with equal probability. For these assumptions, it is interesting to discover that the resulting state equations, though just as numerous as before, can be solved explicitly. Harold Gumbel (10) writes a system of state equations for this problem, assuming that customers who find all servers busy wait until served, and he obtains an explicit solution for all of the state probabilities. He then calculates the utilization of each server, and the error which would result from substituting a homogeneous system, each of whose servers have rates equal to the average of those in the heterogeneous system. Ancker and Gafarian (1) extend Gumbel's work by considering situations where customers leave the system if the waiting line is too long or if not served during an exponentially distributed time interval.

Another solution of this modified problem is given by Kolifrath and Smith (13).

Although their results are essentially the same as those obtained by Gumbel, their methodology involves the calculation of the conditional probabilities that a particular server is busy given that a specified total number is busy. Finally, Krishnamoorthi (14) considers a heterogeneous two-server system with the assumption that if an arrival finds both servers idle he occupies server #1 with probability p and server #2 with probability $1 - p$, where server #1 has the shortest mean service time. An interesting conclusion in this work is that the expected queue length is minimized if p equals 1. Thus, in the case where server #1 is faster than server #2, it is better to direct a customer to the faster server ($p = 1$) than to let the customer choose his server at random ($p = \frac{1}{2}$). Therefore, it seems reasonable to make the assumption of ordered hunt since the strategy implied by that assumption is an optimal one, at least for a system with two heterogeneous servers.

Homogeneous Service Times and Ordered Hunt

The literature on this problem is much more extensive, and so this discussion will be limited to those results which are most relevant to the objective of calculating the utilization of each server. One approach, which is discussed by Cooper (6, pp. 122-124) uses the concept of carried load, which is defined as the average number of busy servers. For a one-server system the carried load is then equal to the utilization of that server. It is argued that, for the case of blocked customers cleared, the load carried by the system comprised of the first m servers ($m \leq s$) is not affected by the existence of higher-numbered servers. Therefore, the carried loads are additive in the sense that the utilization of the m^{th} server is

equal to the difference between the loads carried by the first m and $m - 1$ server groups respectively. Finally, it is demonstrated that these total carried loads can be calculated easily. Paul J. Burke [1963, unpublished; see Cooper (6, p. p. 124)] and Vulot (23) are each credited with extending this result to systems where queueing is allowed.

C. Palm (15), as also discussed by Khintchine (12), considers those arrivals which overflow (are blocked on) the first m servers as comprising the arrival stream to the $(m + 1)^{\text{st}}$ server. Furthermore, if the arrival stream to the m^{th} server is recurrent, or in other words possesses independent, identically distributed interarrival times, then the stream entering the $(m+1)^{\text{st}}$ server is also recurrent. Then, a recurrence relation is given which permits calculation of the interarrival time distribution function for the $(m+1)^{\text{st}}$ server using the same function for the m^{th} server. Hence, this information can be calculated for all servers providing the first server's interarrival time distribution function, which need not be negative exponential, is known. Also, further analysis of each server is made possible since the arrival stream and service time are both adequately specified. Khintchine extends his discussion of Palm's work for the case of Poisson input by calculating an explicit expression for the interarrival time distribution function for the m^{th} server, which turns out to be a mixture of m different exponential functions.

Some later work by A. Descloux (9) is related to that of Palm, but by assuming that the first server arrival stream is Poisson, he is able to simplify the aforementioned recurrence relation for the $(m+1)^{\text{st}}$ server's interarrival time distribution function. Also, the moments of this distribution are calculated.

The difficulty connected with extending Palm's work to the assumption of heterogeneous service times is centered primarily on the calculation of an explicit expression for the $(m+1)^{\text{st}}$ server's interarrival time distribution function.

It will be argued, however, that the inequality of service time means has no effect on Palm's recurrence relation as long as all service times are exponentially distributed. Unfortunately, Descloux's work is not as easily generalized.

CHAPTER III

STATE EQUATION SOLUTION METHOD FOR BLOCKED CUSTOMERS CLEARED CASE

In this chapter, the objective will be to calculate the probability that all s servers are busy, $B(s)$, and the utilization of each server, $p_m \forall$ (for all) $m = 1, 2, \dots, s$, for a heterogeneous system with Poisson arrivals and blocked customers cleared. Traditional methods of defining a state space and writing equations relating the state probabilities will be used.

Development of State Space

Since Poisson arrivals are characterized by independent negative exponentially distributed interarrival times, and since the service times have the same property, the following characteristic of any negative exponential random variable, T , applies to both:

$$\Pr \{T > t + h \mid T > t\} = \Pr \{T > h\} \quad \forall t \geq 0, h \geq 0.$$

Therefore, this queueing system fits the definition of a Markov process in the sense that its probabilistic evolution after time t depends only upon the state of the system at t and is independent of the history of the system prior to t .

For the case where the mean service completion rates can be different, however, the future evolution of the system after time t depends not only upon the

total number of busy servers at t but also upon which of the s servers are busy.

Therefore, it is necessary to define the state space S as follows:

$$S = \{ (X_1, \dots, X_m, \dots, X_s) : X_m \in \{0, 1\} \}$$

$$X_m = \begin{cases} 1 & \text{if server } m \text{ is busy} \\ 0 & \text{otherwise} \end{cases}$$

Thus, each state is represented by an s -dimensional vector of zeros and ones, and there is a total of 2^s possible states. Also, since the number of states is finite, and each state can be reached from every other state at some future time, there exists a unique 2^s -dimensional row vector of stationary state probabilities, P , such that

$$P_{x_1, \dots, x_m, \dots, x_s} = \Pr\{X_1 = x_1, \dots, X_2 = x_2, \dots, X_s = x_s\}.$$

Matrix State Equations for Calculating the Stationary State Distribution

In simplicity of notation it is now desirable to place the states into one to one correspondence with the following space:

$$S' = \{i : i = 1, 2, \dots, s\}.$$

Therefore,

$$P_{x_1, \dots, x_m, \dots, x_s} = P_i$$

for exactly one value of i , and the row vector P remains unchanged.

Now, from work by Spitzer (19), there exists for this queueing system a

unique semigroup of transition matrices $\{Q_t\}_{t \geq 0}$ ($2^S \times 2^S$) whose entries, $Q_t(i, j)$, denote the conditional probability that the system is in state j at time t given that it was in state i at time zero $\forall i, j$ in S' . It follows that for P to be a stationary distribution, it must satisfy the equation

$$P Q_t = P \quad \forall t \geq 0. \quad (3.1)$$

However, rather than trying to solve equation 3.1 for P , it is easier to use what Spitzer calls the transition matrix generator, G , for the matrices $\{Q_t\}$. The generator is defined such that it satisfies the following properties:

1. $G(i, j) \geq 0 \quad \forall i, j$ in S' such that $i \neq j$
2. $\sum_{j=1}^{2^S} G(i, j) = 0 \quad \forall i$ in S'
3. $G(i, j) = Q'_0(i, j) \quad \forall i, j$ in S'

A more intuitively satisfying explanation of the elements of G is that, for $i \neq j$ and $\Delta t \rightarrow 0$, $G(i, j) \Delta t$ is asymptotically equal to the probability that the system is in state j at $t + \Delta t$ given that it was in state i at t . Therefore, $G(i, j)$ can be viewed as a transition rate from state i to state j . The diagonal elements of G are determined from property 2.

The usefulness of the generator stems from the fact that the entries of G are easier to specify than those of $\{Q_t\}$. Also, P is a stationary distribution if and only if it satisfies the following equations:

$$PG = \underline{0} , \quad (3.2)$$

$$\sum_{j=1}^{2^S} P_j = 1 , \quad (3.3)$$

where $\underline{0}$ is a 2^S -dimensional row vector of zeros. Therefore, this problem reduces to one of specifying the entries of G and then finding the unique probability distribution, P , satisfying equations 3.2 and 3.3. Then $B(s)$ and p_m can be found from the following relations:

$$B(s) = P_{1,1,\dots,1} \quad (3.4)$$

$$p_m = \sum_{\{(x_1, \dots, x_m, \dots, x_s) : x_m = 1\}} P_{x_1, \dots, x_m, \dots, x_s} \quad m = 1, 2, \dots, s . \quad (3.5)$$

In order to better illustrate these state equation solution methods and the construction of the matrix generator, G , an explicit solution for the two-server problem follows.

Explicit Solution for the Two-Server Problem

When there are two servers with different mean service completion rates and the BCC queue discipline is assumed, the state space S reduces to

$$\begin{aligned} S &= \{(X_1, X_2) : X_m \in \{0, 1\}\} \\ &= \{(0, 0), (1, 0), (0, 1), (1, 1)\} . \end{aligned}$$

Therefore, the generator, G , is a 4×4 matrix. The procedure for filling out the entries of G is to complete the off-diagonal entries for each row and then use property #2 of the matrix generator to obtain the diagonal element. Considering the first row of G , for $\Delta t \rightarrow 0$ the probability of being in state $(1, 0)$ at time $t + \Delta t$ given that the system was in state $(0, 0)$ at time t is asymptotically equal to $\lambda \Delta t$, and the transition rate from $(0, 0)$ to $(1, 0)$ is λ , the mean arrival rate. For the system to be in states $(0, 1)$ or $(1, 1)$ at time $t + \Delta t$ given in state $(0, 0)$ at t , at least two arrivals must occur in the interval Δt . For Poisson arrivals, the probability of this event is small enough to be ignored, and as a result the corresponding transition rates equal zero. Finally, from property #2, the diagonal entry in the first row of G is $-\lambda$. Continuing this procedure for the other three rows results in the following generator.

$$G = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \mu_1 & -\lambda - \mu_1 & 0 & \lambda \\ \mu_2 & 0 & -\lambda - \mu_2 & \lambda \\ 0 & \mu_2 & \mu_1 & -\mu_1 - \mu_2 \end{bmatrix}$$

Now, to obtain the stationary distribution for this 2 server system we need to solve equation 3.2 where G is as above and

$$P = (P_{0,0}, P_{1,0}, P_{0,1}, P_{1,1}) .$$

The solution for P in this equation can be expressed, after some straightforward

but rather laborious calculations, as follows:

$$P_{1,1} = 1 \cdot P_{1,1}$$

$$P_{0,1} = \frac{\mu_1}{\lambda + \mu_2} P_{1,1}$$

$$P_{1,0} = \frac{\mu_2}{\lambda} \left(1 + \frac{\mu_1}{\lambda + \mu_2}\right) P_{1,1}$$

$$P_{0,0} = \frac{\mu_1 \mu_2}{\lambda^2} \left(1 + \frac{\lambda + \mu_1}{\lambda + \mu_2}\right) P_{1,1}$$

Then from equation 2.3

$$P_{0,0} + P_{1,0} + P_{0,1} + P_{1,1} = 1,$$

so that $P_{1,1}$ is shown to be the following:

$$P_{1,1} = \left[1 + \frac{\mu_1}{\lambda + \mu_2} + \frac{\mu_2}{\lambda} \left(1 + \frac{\mu_1}{\lambda + \mu_2}\right) + \frac{\mu_1 \mu_2}{\lambda^2} \left(1 + \frac{\lambda + \mu_1}{\lambda + \mu_2}\right)\right]^{-1}. \quad (3.7)$$

Therefore, all stationary probabilities have been found and from equations 3.4 and 3.5,

$$\begin{aligned} P_1 &= P_{1,0} + P_{1,1} \quad , \\ P_2 &= P_{0,1} + P_{1,1} \quad , \\ B(2) &= P_{1,1} \quad . \end{aligned} \quad (3.8)$$

The main point to be made from this example was a demonstration of how

the G matrix is formed, and how it leads to the desired results for this queueing problem. However, it can also be seen that the algebra was surprisingly involved, even for the 2-server case. For the 3-server case, G would be an 8 x 8 matrix, which would result in much greater computational difficulty.

Conclusions

It is evident that, using equations 3.2 and 3.3, numerical solutions for p can be found in principle, even for large values of s . Since solving the above matrix equation, however, is equivalent to solving 2^s simultaneous equations in 2^s unknowns, even the problem of obtaining numerical solutions becomes unmanageable as s increases. Furthermore, a usable algebraic solution for each of the entries in P has not as yet been found. Therefore, this method for calculating p_m and $B(s)$ seems useful only for small s . Another point regarding this solution method is that it involves the calculation of state probabilities which are superfluous in that they are not needed in equations 2.3 and 2.4.

CHAPTER IV

ALTERNATIVE SOLUTION FOR BLOCKED CUSTOMERS

CLEARED CASE

Development of Alternate Solution Methods

In this chapter, an alternative method is developed for calculating the blocking probability, $B(s)$ and the utilization of each server, p_m , $m = 1, 2, \dots, s$, for the BCC system with heterogeneous service times. This method involves the adaptation of Palm's work as previously described to the case of heterogeneous service times, and as a result the assumption of Poisson input can be relaxed to that of recurrent input. Also, the first objective will be to look at the system only at the instants immediately preceding customer arrivals, and to calculate the following probabilities:

$$\pi_m = \Pr \{ \text{customer finds } m^{\text{th}} \text{ server busy} \mid \text{he finds all of the first } m-1 \text{ servers busy} \}$$

$$A(m) = \Pr \{ \text{customer finds all of the first } m \text{ servers busy} \}$$

It is noted here, from the definition of conditional probability, that

$$A(m) = \pi_m A(m-1)$$

$$m = 1, 2, \dots, s$$

$$A(0) = 1$$

$$A(m) = \prod_{i=1}^m \pi_i, \quad (4.1)$$

so the determination of $\pi_m \forall m = 1, 2, \dots, s$ leads easily to the calculation of all $A(m)$ as well. In fact, this arriving customer blocking probability, $A(m)$ is of as much interest as $B(m)$, but on the other hand π_m is not a measure of the utilization of the m^{th} server. Therefore, a method will be given for calculating p_m from π_m .

In order to calculate π_m it is best to consider the m^{th} server as a one-server system having negative exponential service times with mean $1/\mu_m$, a BCC queue discipline, and an arrival stream with mean rate λ_m made up of those customers who find all of the first $m - 1$ servers busy. The only difference between this system and the simplest queueing system (one server, BCC, Poisson arrivals, and negative exponential service times) is the stream of arrivals which is not Poisson for $m > 1$. Therefore, the calculation of π_m should depend only upon our ability to characterize the random variable denoting the time between arrivals to the m^{th} server by a cumulative distribution function, $G_m(t)$.

Before attempting to devise a means of calculating $G_m(t)$, it is necessary to establish that the stream of arrivals to the m^{th} server is recurrent. To this end it has been proved by Palm (15) for homogeneous negative exponential service times that if the arrival stream to any server, m , is recurrent, then the arrival stream for the $(m + 1)^{\text{st}}$ server is also recurrent. Therefore, if the input to the system is recurrent, then each of the s servers possesses a recurrent arrival stream. To extend this result to the case of heterogeneous service times, it is merely necessary to observe that, as proved for homogeneous service times, the existence of recurrent input for the $(m + 1)^{\text{st}}$ server requires only that (1) the m^{th} server possesses recurrent input and (2) the service times for the m^{th} server alone be negative

exponentially distributed. There exist no requirements regarding the service time distributions of other servers. Therefore, it is concluded that for the heterogeneous system the entering stream for each of the s servers is recurrent.

Now, in order to calculate $G_m(t)$, the recursive methods described in Syski (20, p. 262) [see also Khintchine (12) and Riordan (16)] will be used. First, we define $G_m^c(t)$ as the probability that no arrival to the m^{th} server occurs in time t . Then, it follows that

$$G_m(t) = 1 - G_m^c(t) \quad (4.2)$$

The objective is to derive an expression for $G_{m+1}^c(t)$ in terms of $G_m^c(t)$ and the service time distribution function of server m , and then to determine $G_{m+1}(t)$ from equation 4.2.

The event that no arrival to the $m+1^{\text{st}}$ server occurs in time t can be partitioned into the following disjoint sub-events:

1. Event E_1 is the event that server m experiences no arrivals in time t , and

$$P_r \{E_1\} = G_m^c(t)$$

2. Event E_2 is the event that at least one arrival to the m^{th} server occurs in time t but none of these arrivals finds the m^{th} server busy.

To calculate $Pr \{E_2\}$, assume that the first arrival to server m occurs in the time interval $(\xi, \xi + d\xi)$. The differential of the probability of an arrival in that interval is $-dG_m^c(\xi)$. Now, for E_2 to occur two other events must also occur. First, we

note that for an arrival to server $m+1$ to have occurred at time zero, server m must have been occupied. It is necessary that this service time be completed by time ξ , which occurs with probability

$$1 - e^{-\mu_m \xi}.$$

Second, there must be no arrivals to the $(m+1)^{\text{st}}$ server in the time interval (ξ, t) , which occurs with probability $G_{m+1}^c(t - \xi)$. This statement of probability, however, needs justification since no arrival to server $m + 1$ occurred at time ξ , and the use of $G_{m+1}^c(t - \xi)$ is predicated on such an arrival.

Consider the first m servers immediately after the arrival at ξ . Since service times are negative exponentially distributed, the service times remaining for each of these m customers have the same set of probability distributions whether or not the arrival at time ξ was served by server m . Also, the occurrence of future arrivals to the system are independent of the disposition of the arrival at time ξ . Therefore, by assuming that the arrival at time ξ was blocked on server m , we have not changed either of the two factors affecting the future evolution of the system, and the use of $G_{m+1}^c(t - \xi)$ in this instance is justified.

Now the above three events comprising E_2 represent realizations of two successive interarrival times and a service time which are mutually independent random variables. Therefore, the probability of their joint occurrence is the product of the individual probabilities, or

$$-(1 - e^{-\mu_m \xi}) G_{m+1}^c(t - \xi) dG_m^c(\xi).$$

Finally, since the first arrival to server m can occur at any time in the interval $(0, t)$, the probability of event E_2 is as follows:

$$\Pr\{E_2\} = - \int_0^t (1 - e^{-\mu_m \xi}) G_{m+1}^c(t - \xi) dG_m^c(\xi)$$

Therefore,

$$\begin{aligned} G_{m+1}^c(t) &= \Pr\{E_1\} + \Pr\{E_2\} \\ &= G_m^c(t) - \int_0^t (1 - e^{-\mu_m \xi}) G_{m+1}^c(t - \xi) dG_m^c(\xi), \end{aligned} \quad (4.3)$$

and a recursion formula has been found which makes possible the calculation of $G_m(t)$, $\forall m = 1, 2, \dots, s$. In order to facilitate these computations, it is best to use the Laplace-Stieltjes transform of $G_m(t)$:

$$\Psi_m(z) = \int_0^\infty e^{-zt} dG_m(t) \quad \text{Re } z \geq 0.$$

Then from equations 4.2 and 4.3, the following recursion formula can be derived in terms of the transforms:

$$\Psi_{m+1}(z) = \frac{\Psi_m(z + \mu_m)}{1 - \Psi_m(z) + \Psi_m(z + \mu_m)} \quad (4.4)$$

Another quantity of interest is the m^{th} server's mean interarrival time, $1/\lambda_m$, and it is noted from a property of these transforms that

$$\frac{1}{\lambda_m} = \frac{d}{dz} \Psi_m(z) \Big|_{z=0}$$

Therefore, by evaluating the derivative with respect to z of both sides of 4.4 and setting z equal to zero, it is found that

$$\begin{aligned} \lambda_{m+1} &= \lambda_m \Psi_m(\mu_m) , \\ &= \lambda \prod_{i=1}^m \Psi_i(\mu_i) \end{aligned} \quad (4.5)$$

where λ equals λ_1 , the mean arrival rate to the first server.

Finally, it can easily be shown that the following is true:

$$\pi_m = \Psi_m(\mu_m) \quad (4.6)$$

So show this, let time zero be an arrival instant for server m . Then, the differential of the probability that the next arrival to server m occurs in the interval $(t, t + dt)$ is $dG_m(t)$. For this customer to find the m^{th} server busy, the service time which either began or was already in progress at time zero must still be in progress, which occurs with probability $e^{-\mu_m t}$. Therefore, the probability that a customer arrives to server m at time t and finds the server busy is

$$e^{-\mu_m t} dG_m(t)$$

Finally, since t can be anywhere in the interval $[0, \infty)$, the probability that an arrival to the m^{th} server finds it busy is

$$\begin{aligned}\pi_m &= \int_0^{\infty} e^{-\mu_m t} dG_m(t) , \\ &= \Psi_m(\mu_m) ,\end{aligned}$$

thus proving equation 4.6.

As previously stated, π_m does not equal the utilization of the m^{th} server. This is true since the stream of arrivals at the m^{th} server ($m > 1$) is not a Poisson stream. Stated another way, π_m is conditioned on the event that the arriving customer finds all of the first $m-1$ servers busy, and the state probabilities for the m^{th} server are affected by such knowledge of the states of the first $m-1$ servers.

Therefore, the conditional probability, π_m , is not the same as the unconditional probability, p_m . To calculate p_m , the work of Takács (21, pp. 182-185) is used. He states that for an s -server queueing system with homogeneous negative exponential service times, recurrent input, and blocked customers cleared, that the stationary state distribution for any arbitrary instant can be calculated easily if the arriving customer's state distribution is known. He then gives the formula for making this computation.

These results can be applied to the system with heterogeneous service times and ordered hunt since each server in the group is a one-server system fulfilling Takács' conditions. His formula, applied to the m^{th} server, is

$$p_m = \frac{\lambda_m}{\mu_m} (1 - \pi_m). \quad (4.7)$$

Finally, using equations 4.5 and 4.6, equations 4.1 and 4.7 can be restated as follows:

$$A(m) = \prod_{i=1}^m \Psi_i(\mu_i), \quad m = 1, 2, \dots, s \quad (4.8)$$

$$p_m = \begin{cases} \frac{\lambda}{\mu_m} [1 - \Psi_m(\mu_m)], & m = 1 \\ \frac{\lambda}{\mu_m} \left[\prod_{i=1}^{m-1} \Psi_i(\mu_i) \right] [1 - \Psi_m(\mu_m)], & m = 2, 3, \dots, s \end{cases} \quad (4.9)$$

It is noted from a discussion in Cooper (6, p. 65) that if the input to the first server is Poisson, then the fact that the system is at an arrival instant, as opposed to an arbitrary instant, adds nothing to our knowledge of the times of previous arrivals or service completions, which are the only factors determining the present state of the system. Thus, the state probabilities are also unaffected and the following is true:

$$A(m) = B(m) \quad m = 1, 2, \dots, s. \quad (4.10)$$

However, since p_m and $A(m)$ satisfy the objectives stated earlier, the assumption of Poisson input is not needed for the blocked customers cleared case. Nevertheless, in order to extend these results to the case where queueing is allowed, as is done in the next chapter, the assumption of Poisson input is needed.

Assessment of the Efficiency of These Computations

In terms of using these methods for calculating p_m and $A(m)$, it is noted

that closed-form (non-recursive) expressions for $G_m(t)$ and $\Psi_m(z)$ have not been found. Such closed-form expressions are complicated even for Poisson input and homogeneous negative exponential service times, as was stated previously. Fortunately, all that is required for these calculations are numerical evaluations of $\Psi_m(\mu_m)$ $\Psi_m = 1, 2, \dots, s$. In order to determine the number of calculations required it is best to look at $\Psi_s(\mu_s)$ and determine from equation 3.4 the needed values of $\Psi_{s-1}(z)$. Then this process is repeated for each of these values of $\Psi_{s-1}(z)$ to find the necessary values of $\Psi_{s-2}(z)$. These iterations are continued until all needed values of $\Psi_1(z)$ have been determined. The result is that $\Psi_1(z)$ must be evaluated at $2^s - 1$ different arguments, and then $2^s - s - 1$ iterations on equation 4.4 must be performed. An example of this procedure for $s = 4$, is shown on Exhibit I on page 27.

Another Method of Calculating the Utilization of Each Server

In order to check the results of this chapter, another method was developed for calculating p_m . Rather than calculating π_m and then employing Takács' theorem, this method argues that

$$p_m = \frac{\tau_m}{E\{\text{cycle length}\}} \quad , \quad (4.11)$$

where the cycle length is the time between transitions from an idle to a busy state, and τ_m is the mean service time ($1/\mu_m$). Hence we must prove equation 4.11 and then calculate the expected cycle length. This is done in the appendix, and the results agree with equation 4.9.

Alternate Solution for Two-Server Problem

As an example of how the methods of this chapter should be used, the two-server problem is solved as follows. First, since the input to the first server is Poisson,

$$G_1(t) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then it can be easily shown that

$$\Psi_1(z) = \frac{\lambda}{\lambda + z}, \quad \text{Re } z \geq 0.$$

In order to apply the alternate method the following two evaluations of $\Psi_1(z)$ are required:

$$\Psi_1(\mu_1) = \frac{\lambda}{\lambda + \mu_1}$$

$$\Psi_1(\mu_1 + \mu_2) = \frac{\lambda}{\lambda + \mu_1 + \mu_2}.$$

Then, using equation 4.4 and evaluating $\Psi_2(z)$ at μ_2

$$\begin{aligned} \Psi_2(\mu_2) &= \frac{\Psi_1(\mu_1 + \mu_2)}{1 - \Psi_1(\mu_2) + \Psi_1(\mu_1 + \mu_2)}, \\ &= \frac{\frac{\lambda}{\lambda + \mu_1 + \mu_2}}{1 - \frac{\lambda}{\lambda + \mu_2} + \frac{\lambda}{\lambda + \mu_1 + \mu_2}} \end{aligned}$$

Finally, from the assumption of Poisson input it is true that

$$A(2) = B(2)$$

Therefore, from equations 4.8 and 4.10, the probability that both servers are busy is

$$\begin{aligned} B(2) &= \Psi_1(\mu_1) \Psi_2(\mu_2) , \\ &= \left[\frac{\lambda}{\lambda + \mu_1} \right] \left[\frac{\frac{\lambda}{\lambda + \mu_1 + \mu_2}}{1 - \frac{\lambda}{\lambda + \mu_2} + \frac{\lambda}{\lambda + \mu_1 + \mu_2}} \right] . \end{aligned} \quad (4.12)$$

From equation 4.9

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu_1} \left[1 - \frac{\lambda}{\lambda + \mu_1} \right] , \\ &= \frac{\lambda}{\lambda + \mu_1} , \end{aligned}$$

and

$$p_2 = \frac{\lambda}{\mu_2} \left[\frac{\lambda}{\lambda + \mu_1} \right] \left[1 - \frac{\frac{\lambda}{\lambda + \mu_1 + \mu_2}}{1 - \frac{\lambda}{\lambda + \mu_2} + \frac{\lambda}{\lambda + \mu_1 + \mu_2}} \right] . \quad (4.14)$$

It can be demonstrated that these results are equivalent to those in equations 3.6, 3.7, and 3.8. Of course, the actual use of this alternate method for a larger number of servers would not require that such equations as 4.12 and 4.14 be written out explicitly.

Exhibit I. Calculating $\Psi_m(\mu_m)$ For $s = 4$

Server Number m	Number of Iterations on Eq. 4.4	Needed values of $\Psi_m(z)$	Needed Values of $\Psi_{m-1}(z)$ for Use of Eq. 4.4
4	1	$\Psi_4(\mu_4)$	$\Psi_3(\mu_4), \Psi_3(\mu_3 + \mu_4)$
3	3	$\Psi_3(\mu_3)$	$\Psi_2(\mu_3), \Psi_2(\mu_2 + \mu_3)$
		$\Psi_3(\mu_4)$	$\Psi_2(\mu_4), \Psi_2(\mu_2 + \mu_4)$
		$\Psi_3(\mu_3 + \mu_4)$	$\Psi_2(\mu_3 + \mu_4), \Psi_2(\mu_2 + \mu_3 + \mu_4)$
2	7	$\Psi_2(\mu_2)$	$\Psi_1(\mu_2), \Psi_1(\mu_1 + \mu_2)$
		$\Psi_2(\mu_3)$	$\Psi_1(\mu_3), \Psi_1(\mu_1 + \mu_3)$
		$\Psi_2(\mu_2 + \mu_3)$	$\Psi_1(\mu_2 + \mu_3), \Psi_1(\mu_1 + \mu_2 + \mu_3)$
		$\Psi_2(\mu_4)$	$\Psi_1(\mu_4), \Psi_1(\mu_1 + \mu_4)$
		$\Psi_2(\mu_2 + \mu_4)$	$\Psi_1(\mu_2 + \mu_4), \Psi_1(\mu_1 + \mu_2 + \mu_4)$
		$\Psi_2(\mu_3 + \mu_4)$	$\Psi_1(\mu_3 + \mu_4), \Psi_1(\mu_1 + \mu_3 + \mu_4)$
		$\Psi_2(\mu_2 + \mu_3 + \mu_4)$	$\Psi_1(\mu_2 + \mu_3 + \mu_4), \Psi_1(\mu_1 + \mu_2 + \mu_3 + \mu_4)$
1	0	$\Psi_1(\mu_1)$	
	$\Psi_1(z)$ is known	$\Psi_1(\mu_2)$	
		$\Psi_1(\mu_1 + \mu_2)$	
		$\Psi_1(\mu_3)$	
		$\Psi_1(\mu_1 + \mu_3)$	
		$\Psi_1(\mu_2 + \mu_3)$	
		$\Psi_1(\mu_1 + \mu_2 + \mu_3)$	
		$\Psi_1(\mu_4)$	
			$15 = 2^4 - 1$ Evaluations of $\Psi_1(z)$

Exhibit I. Calculating $\Psi_m(\mu_m)$ For $s = 4$ (Continued)

Server Number m	Number of Iterations on Eq. 4.4	Needed Values of $\Psi_m(z)$	Needed Values of $\Psi_{m-1}(z)$ For Use of Eq. 4.4
1 (continued)	0 (continued)	$\Psi_1(\mu_1 + \mu_4)$ $\Psi_1(\mu_2 + \mu_4)$ $\Psi_1(\mu_1 + \mu_2 + \mu_4)$ $\Psi_1(\mu_3 + \mu_4)$ $\Psi_1(\mu_1 + \mu_3 + \mu_4)$ $\Psi_1(\mu_2 + \mu_3 + \mu_4)$ $\Psi_1(\mu_1 + \mu_2 + \mu_3 + \mu_4)$	
Total	11 = $2^4 - 4 - 1$		

CHAPTER V

EXTENSION OF RESULTS TO A SYSTEM WITH QUEUEING ALLOWED

Definitions and Assumptions

Having succeeded in calculating $B(s)$ and p_m ($m = 1, 2, \dots, s$) for the case of heterogeneous service times, ordered hunt, and blocked customers cleared, it is the objective now to extend these results to the comparable system where a queue is allowed to form. Furthermore, we wish to consider all types of queue disciplines for which the system can continue to be described as a Markov process. Therefore, in addition to Poisson arrivals, we need only assume that when all servers are busy and there are k customers waiting to be served, the time, T , to the next service completion or defection from the queue (whichever comes first) follows the negative exponential probability distribution with mean $1/\beta_k$:

$$P(T < t) = 1 - e^{-\beta_k t} \quad t > 0$$

$$k = 0, 1, 2, \dots$$

Notice that no assumption is made regarding the order of service of waiting customers.

More specifically, the following two assumptions fulfill the above requirements:

1. Customers who find all s servers busy join the queue and then wait until served.

For this case, if $\mu = \sum_{m=1}^s \mu_m$,

$$\beta_k = \mu \quad k = 0, 1, 2, \dots$$

This is commonly called a BCD (blocked customers delayed) queue discipline, and specific results for this assumption will be given later in this chapter.

2. Customers who find all servers busy join the queue and remain there either until served or until their waiting time exceeds a random interval having the negative exponential distribution with parameter α . For this case

$$\beta_k = \mu + k\alpha \quad k = 0, 1, 2, \dots$$

In order to redefine our objectives in terms of this generalized assumption, the following state space definition is required:

$$S = \{(X_1, \dots, X_m, \dots, X_s, k) : X_m \in \{0, 1\}, k = 0, 1, 2, \dots\}$$

where $X_m = \begin{cases} 1 & \text{if server } m \text{ is busy,} \\ 0 & \text{if server } m \text{ is idle,} \end{cases}$

and $k = \text{number of customers in the queue.}$

Also, $\hat{P}_{x_1, \dots, x_s, k} = \Pr \{X_1 = x_1, \dots, X_s = x_s, k \text{ customers waiting}\}.$

Whereas in the BCC case we wanted to find $B(s)$, the probability that all servers are busy, here the analogous quantity is:

$$C(s) = \sum_{k=0}^{\infty} \hat{P}_{1, 1, \dots, 1, k}$$

Whereas in the BCC case we wanted to find p_m for $m = 1, 2, \dots, s$, the utilization of each server, here the same quantity is:

$$\hat{p}_m = \sum_{\{(x_1, \dots, x_m, \dots, x_s, k) : x_m = 1\}} \hat{P}_{x_1, \dots, x_m, \dots, x_s, k}$$

Here, as in the BCC case, a solution is desired which doesn't depend upon the solution of state equations involving these probabilities.

It appears that the greatest hope for a straight-forward generalization of the BCC case lies in the possibility that the relationships between p_m and \hat{p}_m and also $B(s)$ and $C(s)$ are simple ones. Since these quantities are all sums of stationary state probabilities, which are in turn determined by the matrix generator, a comparison of G for a system with BCC and \hat{G} for the corresponding system where queueing is allowed should help determine whether or not such simple relationships exist. To make such a comparison, it is desirable to consider matrices of finite dimension. Therefore, we assume for the generalized case that there are n waiting positions and that arrivals finding all servers busy and n other customers waiting in line are cleared from the system. Then, if we consider only those systems for which $P_{1, \dots, 1, k}$ approaches zero as k approaches infinity n can be chosen large enough so that the probability that customers find all waiting positions occupied is essentially zero. In that case, the stationary state probabilities will be the same as those for a system with an infinite number of waiting positions.

Construction of Transition Matrix Generator

With these assumptions in mind, the matrix generator \hat{G} for the general

case will be of dimension $(2^S + n) \times (2^S + n)$. To demonstrate its construction, we partition \hat{G} as follows:

$$\hat{G} = \begin{array}{c|c} \begin{array}{c} 2^S \times 2^S \\ A \end{array} & \begin{array}{c} 2^S \times n \\ B \end{array} \\ \hline \begin{array}{c} n \times 2^S \\ C \end{array} & \begin{array}{c} n \times n \\ D \end{array} \end{array}$$

The matrix A contains the transition rates from those 2^S states where $k = 0$ to that same set of states. It will be argued that, except for the last diagonal element (for state $(1, \dots, 1, 0)$), the entries of A are exactly the same as those of G, the generator for the corresponding BCC queueing system. This is seen to be true from the following:

1. The transition rates to and from states in which no customers are waiting are the same whether or not queueing is allowed. Therefore, the non-diagonal elements of A are the same as those of G.
2. The matrix B contains transition rates from states where there are no customers waiting to states where there is a queue. However, with a single exception, it is not possible to make such transitions in one step. The one exception is the transition from $(1, \dots, 1, 0)$ to $(1, \dots, 1, 1)$, which is made with rate λ . Therefore, matrix B has a lower left entry equal to λ and all other entries equal to zero. Also, since each diagonal entry in A equals minus the sum of the off-diagonal entries in its row, all of the diagonal entries of A, except for the last, will be the same as those of G. The last diagonal entry of A will be $G(2^S, 2^S) - \lambda$.

We have noted the construction of matrix B, and now argue that the C matrix also has zeros everywhere except for the upper right hand element, which is the transition rate, β_1 , from the state with all servers busy and one customer waiting in the queue to the state with all servers busy and no customers waiting.

Finally, matrix D has the transition rates to and from those states where a queue exists. Notice that, for each of these n states, all servers must be busy. Therefore, the only index which varies is k , and the problem reduces to one of only one dimension:

$$\hat{P}_{1, \dots, 1, k} \rightarrow \hat{P}_k \quad k = 0, 1, \dots, n \quad (5.1)$$

For all intermediate states where $2 \leq k \leq n-1$, the entries in D are λ for an upward transition from state k to state $k+1$, β_k for a downward transition from k to $k-1$, and $-\lambda - \beta_k$ on the diagonal. A diagram of \hat{G} is given in Exhibit 2 showing the construction of this matrix in its entirety.

Exhibit II. The Matrix Generator, \hat{G}

$$\hat{G} = \begin{array}{c|cccc} & \text{G} & & & \text{O} \\ & G(2^s, 2^s) - \lambda & & & \\ \hline & \beta_1 & -(\lambda + \beta_1) & \lambda & 0 \dots \dots \dots 0 \\ & & \beta_2 & -(\lambda + \beta_2) & \lambda \quad 0 \dots \dots \dots 0 \\ & \text{O} & 0 & & \vdots \\ & & \vdots & & \vdots \\ & & \vdots & & \vdots \\ & & 0 \dots \dots \dots 0 & \beta_{n-1} & -(\lambda + \beta_{n-1}) \quad \lambda \\ & & 0 \dots \dots \dots 0 & & \beta_n \quad -\beta_n \end{array}$$

Calculation of Results for Generalized Queueing Case

Having calculated the matrix \hat{G} and having shown its relationship to G , the corresponding generator for the BCC case, we now propose the following relationship, between the respective stationary state probabilities:

First of all,

$$\hat{P}_{x_1, \dots, x_s, 0} = c P_{x_1, \dots, x_s} \quad (5.2)$$

for all possible (x_1, \dots, x_s) .

Also

$$\hat{P}_k = \frac{\lambda^k}{\beta_1 \beta_2 \dots \beta_k} \hat{P}_0 \quad k = 1, 2, \dots, n, \quad (5.3)$$

where \hat{P}_k and \hat{P}_0 are as defined in equation 5.1. Equation 5.2 states that each BCC stationary state probability is related to its counterpart for the queueing case by a constant which is the same for all 2^s pairs of probabilities. A proof of this follows. Let H be the matrix comprised of the first 2^s rows and the first $2^s - 1$ columns of \hat{G} . Then H is also the matrix formed by deleting the right hand column of G . Now, since the rows of G all sum to zero, the last column of G is necessarily a linear combination of its other columns. Therefore, any vector, \underline{Y} of 2^s components which is orthogonal to the first $2^s - 1$ columns of G must also be orthogonal to the last column, and the following is true:

$$G' \underline{Y} = \underline{0} \Leftrightarrow H' \underline{Y} = \underline{0}, \quad (5.4)$$

where $\underline{0}$ is any column vector with all entries equal to zero. Next it is remembered from equation 3.2 that the set of vectors which satisfy the equation

$$G' \underline{Y} = \underline{0}$$

is as follows:

$$\{\underline{Y} : \underline{Y} = c \underline{P}', c \text{ real}\},$$

where \underline{P} is again the row vector of 2^S components which contains the probabilities of the unique stationary state distribution for the BCC case. From 5.4, then, we know that the set of solutions to $H' \underline{Y} = \underline{0}$ is exactly the same as above. Finally, noting that the first $2^S - 1$ columns of \hat{G} are of the form,

$$\begin{bmatrix} H \\ \text{---} \\ 0 \end{bmatrix},$$

we see that

$$\begin{bmatrix} H \\ \text{---} \\ 0 \end{bmatrix}' \hat{\underline{P}} = \underline{0}$$

Therefore, it must be true that one member of the set of solutions to $H' \underline{Y} = 0$ must be the column vector formed by deleting the bottom n components of $\hat{\underline{P}}'$. This then proves that equation 5.2 holds. Equation 5.3 gives a set of stationary state probabilities for the situation where all servers are busy. As such, these probabilities must satisfy the matrix equation,

$$\hat{\underline{P}} \hat{G} = \underline{0}',$$

where \hat{P} is the row vector of $2^s + n$ probabilities comprising the stationary state distribution for the generalized system with queueing. From examination of the matrix G , it can be seen that the set of equations which the solutions in 5.3 must satisfy reduces to the following:

$$\lambda \hat{P}_{k-1} - (\lambda + \beta_k) \hat{P}_k + \beta_{k+1} \hat{P}_{k+1} = 0, \quad k = 1, 2, \dots, n-1$$

$$\lambda \hat{P}_{n-1} - \beta_n \hat{P}_n = 0.$$

The fact that the solution form in equation 5.3 is the correct one for these equations is easily verified, and therefore all of these probabilities have been found providing that \hat{P}_0 can be calculated. Also, as long as the series

$$\sum_{k=1}^{\infty} \frac{\lambda^k}{\beta_1 \dots \beta_k}$$

converges, we can rewrite equation 5.3 as follows:

$$\hat{P}_k = \frac{\lambda^k}{\beta_1 \dots \beta_k} \hat{P}_0 \quad k = 0, 1, 2, \dots$$

If this series were to diverge, there would be no stationary state distribution.

Now, it follows from equation 5.2 that

$$\hat{P}_0 = c B(s), \quad (5.5)$$

and that
$$\sum_{\{(x_1, \dots, x_s, k): k=0\}} \hat{P}_{x_1, \dots, x_s, k} = \Pr\{k=0\} = c \quad (5.6)$$

Therefore, using equations 5.5 and 5.6 and the fact that the sum of all state probabilities for the queueing case is one,

$$c + \sum_{k=1}^{\infty} \frac{\lambda^k}{\beta_1 \beta_2 \dots \beta_k} c B(s) = 1 ,$$

or

$$c = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{\beta_1 \beta_2 \dots \beta_k} B(s) \right]^{-1} . \quad (5.7)$$

Now, in order to calculate \hat{p}_m , the event that the m^{th} server is busy can be partitioned into two disjoint sub-events, E_1 and E_2 , such that

$$\hat{p}_m = P_r \{E_1\} + \Pr \{E_2\} , \quad (5.8)$$

where E_1 is the event that the m^{th} server is busy and there is no queue, and E_2 is the event that a queue exists (in which case server m must be busy). Now, it follows from equation 5.2 that

$$\begin{aligned} \Pr \{E_1\} &= \sum_{\{(x_1, \dots, x_m, \dots, x_s, k) : k=0, x_m=1\}} \hat{p}_{x_1, \dots, x_m, \dots, x_s, k} , \\ &= c \sum_{\{(x_1, \dots, x_m, \dots, x_s) : x_m=1\}} p_{x_1, \dots, x_m, \dots, x_s} , \\ &= c p_m \end{aligned}$$

Also, from equation 5.6

$$\begin{aligned} \Pr \{E_2\} &= 1 - \sum_{\{(x_1, \dots, x_s, k): k=0\}} \hat{P}_{x_1, \dots, x_s, k} \\ &= 1 - c. \end{aligned}$$

Thus, from equation 5.8, the following is true:

$$\hat{p}_m = cp_m + (1 - c), \quad (5.9)$$

where c is given by equation 5.7. Also, $C(s)$ can be calculated as follows:

$$C(s) = \hat{P}_0 + \sum_{k=1}^{\infty} \hat{P}_k.$$

Now, from equations 5.5 and 5.6

$$C(s) = cB(s) + (1 - c) \quad (5.10)$$

Finally, it can be seen from equation 5.7, 5.9, and 5.10 that both \hat{p}_m and $C(s)$ have been expressed in terms of quantities which can be calculated using the methods of Chapter IV, and our objective is met.

An Interesting Probabilistic Interpretation

In Chapter II it was mentioned that Burke and Vulot derived a method for extending the analysis to allow queueing, even though the argument might seem at first glance to be valid only for BCC. Their reasoning is discussed in Cooper

(6, pp. 122-124), and it provides an interesting probabilistic interpretation of equations 5.9 and 5.10. Notice that the following can be written from the law of conditional probability and from equations 5.2 and 5.6:

$$\begin{aligned}
 P_{x_1, \dots, x_s} &= \frac{\hat{p}_{x_1, \dots, x_s, 0}}{c} \\
 &= \frac{\Pr\{X_1 = x_1, \dots, X_s = x_s, k = 0\}}{\Pr\{k = 0\}} \\
 &= \Pr\{X_1 = x_1, \dots, X_s = x_s \mid k = 0\} \quad (5.11)
 \end{aligned}$$

Equation 5.11 states that the behavior of the generalized system, given that there is no queue, is exactly the same as the behavior of the corresponding BCC system. In other words, the behavior of the generalized system during time intervals when a queue exists has no effect on the system during subsequent intervals when there is no queue. Therefore, by realizing that equation 5.11 is true because the system has the memoryless properties of a Markov Process, equation 5.9 could have been derived as follows:

$$\begin{aligned}
 \hat{p}_m &= p_m [\Pr\{k = 0\}] + 1 [\Pr\{k > 0\}] \\
 &= c p_m + (1 - c)
 \end{aligned}$$

Also, equation 5.10 could have been derived in a similar fashion. Therefore, these computations depend only on the ability to calculate the probability that at least one customer is waiting. Finally, this reasoning is not restricted to systems with

ordered hunt, but can be extended to any multiserver queueing system which has the properties of a Markov Process.

The Special Case of Blocked Customers Delayed

As a conclusion to the current chapter, we will perform the generalization of our BCC results to the case where blocked customers wait until served (BCD). For this case we stated that when all s servers are busy, the downward transition rates are

$$\beta_k = \mu \quad k = 0, 1, 2, \dots, \quad (5.12)$$

where

$$\mu = \sum_{m=1}^s \mu_m .$$

It was noted that in order to make the assumption of an infinite number of waiting positions and still have a stationary state distribution for the system, it must be true that:

$$\sum_{k=1}^{\infty} \frac{\lambda^k}{\beta_1, \dots, \beta_k} < \infty \quad (5.13)$$

Now let

$$a = \frac{\lambda}{\mu} .$$

Substituting from 5.12 into 5.13, the following requirement is obtained for the BCD case:

$$\sum_{k=1}^{\infty} a^k = \frac{1}{1-a} - 1$$

$$= \frac{a}{1-a}$$

Therefore, from equation 5.7

$$c = \left[1 + \frac{a}{1-a} B(s) \right]^{-1}$$

and equation 5.9 and 5.10 can be used to calculate \hat{p}_m and $C(s)$.

CHAPTER VI

RECOMMENDATIONS FOR FURTHER STUDY

In this work, a new method has been proposed for studying multi-server queueing systems with heterogeneous service times and ordered hunt. It was demonstrated that the proposed approach is computationally appealing relative to the traditional method of writing and solving probability state equations.

One area for further study is to extend the work of Singh, as discussed in Chapter II, to the s -server system. More specifically, assuming that the arrival rate, λ , and sum of the service completion rates, μ , were fixed, find an allocation of service capacity among the s servers which would minimize the blocking probability, $B(s)$, for the blocked customers cleared case. Then show that this allocation will also minimize $E(Q)$ for the generalized case.

Another opportunity for further study lies in increasing the efficiency of the new method by deriving an explicit expression for the m^{th} server's interarrival time distribution function, $G_m(t)$, or its Laplace-Stieltjes transform $\Psi_m(z)$ and thereby eliminating the need for the large number of iterations on equation 4 of Chapter IV. As mentioned in Chapter II, such explicit solutions have been found for homogeneous service times by Palm and others. However, judging from the complexity of their solutions, it is probable that an extension to the case of heterogeneous service times would be difficult. On the other hand, with simplifying assumptions the problem could become more manageable. One such assumption,

which leads to a queueing problem of practical interest, is that there are two server groups, a primary group and an overflow group, whose members have service completion rates μ_1 and μ_2 , respectively. Under this assumption, the main difficulty would lie in determining the interarrival time distribution for the overflow group and the utilization of its servers.

APPENDIX

ALTERNATE METHOD OF CALCULATING THE UTILIZATION OF EACH SERVER

The first step in developing the alternate method of calculating p_m is to show that

$$p_m = \frac{\tau_m}{E\{X_m\}} \quad , \quad (A.1)$$

where X_m is a random variable denoting the time between transitions of server m from an idle to a busy state. Takaćs (21), in his proof of the theorem used in Chapter III, provides two methods for calculating p'_m . One of these methods serves as a proof of A.1, and a summary of this method follows.

Initially, it is demonstrated that the stationary probability p_m exists and is independent of the state of the m^{th} server at time zero. Next, Takaćs defines $N(t)$ as the number of transitions from a busy to an idle state in the time interval $(0, t]$, and $M(t)$ as the number of transitions from an idle to a busy state in the same interval. Also, it is proved that

$$\lim_{t \rightarrow \infty} \frac{M(t)}{t} = \frac{1}{E(X_m)} \quad .$$

Takaćs then argues that if a time variable is introduced for only those periods when the m^{th} server is busy, the transitions from a busy to an idle state form a Poisson

process with mean rate μ_m . Therefore

$$\lim_{t \rightarrow \infty} \frac{N(t)}{t} = \mu_m p_m$$

Finally, it is argued that

$$|M(t) - N(t)| \leq 1 \quad \forall t > 0$$

and therefore that

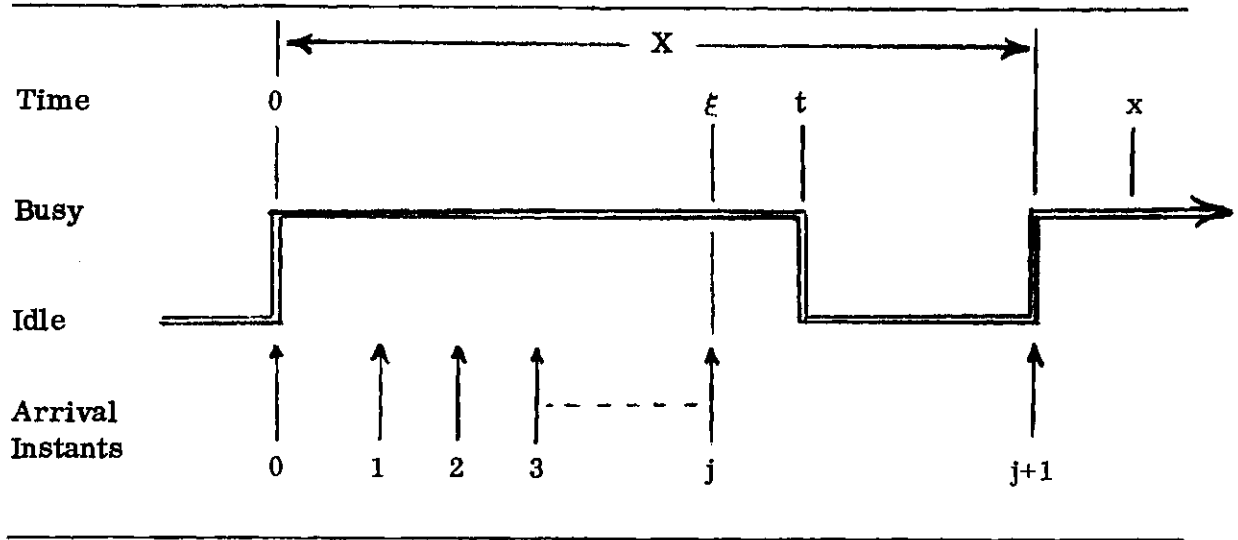
$$\lim_{t \rightarrow \infty} \frac{M(t)}{t} = \lim_{t \rightarrow \infty} \frac{N(t)}{t}$$

from which equation A.1 clearly follows.

Now, the objective is to derive an expression for the cumulative distribution function of the cycle length, $F(x)$, in terms of the service time and interarrival time distribution functions, which are both known. To do this, the probability that a cycle time is less than x must be evaluated. To this end, consider Exhibit III, which is a diagram of an arbitrary cycle with length X . For simplicity of notation, the subscript m has been deleted.

Clearly, for X to be less than x , the termination of the service time beginning at time zero, whose length is denoted by the random variable T , must occur at some time t which is less than x . To begin this analysis, it is desirable to calculate conditional probabilities for x given that T is in the interval $(t, t+dt)$. Then, if t is less than x , X will be less than x if and only if the first arrival after t occurs in the interval (t, x) . The event that the first arrival after t occurs in the

Exhibit III. Diagram of an Arbitrary Cycle of Length X



interval (t, x) can be subdivided into a series of disjoint sub-events according to the number of arrivals occurring in the interval $[0, t)$. Accordingly, let $E_j | t$ represent the event that j arrivals occur in $[0, t)$ and that the $(j+1)^{st}$ occurs in the interval (t, x) . We wish to calculate $P_r \{E_j | t\}$ for all non-negative j . First, for $j = 0$ it is clear that

$$P_r \{E_0 | t\} = G(x) - G(t) . \quad (A.2)$$

For $j > 0$, let ξ be the time of the j^{th} arrival. By definition it must be true that $0 \leq \xi \leq t$. Now, the differential of the probability that the j^{th} arrival occurs in the interval $(\xi, \xi + d\xi)$ is equal to $dG^{*j}(\xi)$, where G^{*j} is the j -fold convolution of G or the cumulative distribution function for the sum of j interarrival times. Next, since the $(j+1)^{st}$ interarrival time is independent of the first j , the probability that the j^{th} arrival is in the interval $(\xi, \xi + d\xi)$ and that the $(j+1)^{st}$ arrival is in the

interval (t, x) is

$$[G(x - \xi) - G(t - \xi)] dG^{*j}(\xi).$$

Finally, since ξ can vary from 0 to t , we have

$$\Pr\{E_j | t\} = \int_0^t [G(x - \xi) - G(t - \xi)] dG^{*j}(\xi) \quad (A.3)$$

and if we define $G^{*0}(\xi) = 1 \quad \forall \quad \xi > 0$, then equation A.3 reduces to equation A.2 for $j = 0$. Therefore, it can be stated that

$$\Pr\{E_j | t\} = \int_0^t [G(x - \xi) - G(t - \xi)] dG^{*j}(\xi) .$$

$$j = 0, 1, 2, \dots$$

Now, since the interarrival times are independent of the service time, $\Pr\{E_j, t\}$, the joint probability that the service time is completed in the interval $(t, t + dt)$, j customers arrive in $[0, t)$, and the $(j + 1)^{st}$ arrives in (t, x) , is as follows:

$$\begin{aligned} \Pr\{E_j, t\} &= P\{E_j | t\} P_r\{t < T < t + dt\} \\ &= \int_0^t [G(x - \xi) - G(t - \xi)] dG^{*j}(\xi) \mu e^{-\mu t} dt \end{aligned} \quad (A.4)$$

Then the marginal probability of the event, E_j , can be found by integrating the right side of equation A.4 with respect to t , where t can be anywhere in the interval $(0, x)$:

$$\Pr\{E_j\} = \int_0^x \int_0^t [G(x-\xi) - G(t-\xi)] dG^{*j}(\xi) \mu e^{-\mu t} dt$$

Finally, from the disjointness of the events E_j , $j = 0, 1, 2, \dots$, the cumulative distribution function for the random variable X can be stated as follows:

$$\begin{aligned} F(x) &= \sum_{j=0}^{\infty} \Pr\{E_j\} \\ &= \sum_{j=0}^{\infty} \int_0^x \int_0^t [G(x-\xi) - G(t-\xi)] dG^{*j}(\xi) \mu e^{-\mu t} dt \end{aligned}$$

Now in order to evaluate $E(X)$ it is convenient to use the Laplace-Stieltjes transform:

$$\gamma(z) = \int_0^{\infty} e^{-zx} dF(x) \quad \text{Re } z \geq 0$$

In order to determine $\gamma(z)$, $\frac{d}{dx} F(x)$ first needs to be evaluated. In order to do this we use a theorem taken from Apostol (2), which states that if

$$F(x) = \int_{p(x)}^{q(x)} f(t, x) dt,$$

and f , q , and p are differentiable with respect to x , then

$$\frac{dF(x)}{dx} = \int_{p(x)}^{q(x)} \frac{d}{dx} f(t, x) dt + f[q(x), x] \frac{dq(x)}{dx} - f[p(x), x] \frac{dp(x)}{dx}. \quad (A.5)$$

First, it is true that

$$\frac{d}{dx} F(x) = \sum_{j=0}^{\infty} \frac{d}{dx} \int_0^x \int_0^t [G(x - \xi) - G(t - \xi)] dG^{*j}(\xi) \mu e^{-\mu t} dt.$$

Then, from equation A.5, let

$$p(x) = 0, \quad \frac{dp(x)}{dx} = 0$$

$$q(x) = x, \quad \frac{dq(x)}{dx} = 1$$

$$\text{and} \quad f(t, x) = \mu e^{-\mu t} \int_0^t [G(x - \xi) - G(t - \xi)] dG^{*j}(\xi).$$

Substituting into equation A.5

$$\begin{aligned} \frac{d}{dx} F(x) &= \sum_{j=0}^{\infty} \left\{ \int_0^x \frac{d}{dx} \left[\mu e^{-\mu t} \int_0^t (G[x - \xi] - G[t - \xi]) dG^{*j}(\xi) \right] dt \right. \\ &\quad \left. + \mu e^{-\mu x} \int_0^x [G(x - \xi) - G(x - \xi)] dG^{*j}(\xi) - 0 \right\} \\ &= \sum_{j=0}^{\infty} \int_0^x \mu e^{-\mu t} \frac{d}{dx} \left[\int_0^t (G[x - \xi] - G[t - \xi]) dG^{*j}(\xi) \right] dt \end{aligned}$$

Now, using the theorem once again, let

$$p(x) = 0, \quad \frac{dp(x)}{dx} = 0$$

$$q(x) = t, \quad \frac{dq(x)}{dx} = 0$$

and
$$f(\xi, x) = [G(x - \xi) - G(t - \xi)] \frac{d}{d\xi} G^{*j}(\xi) .$$

Then

$$\frac{d}{dx} F(x) = \sum_{j=0}^{\infty} \int_0^x \mu e^{-\mu t} \int_0^t \frac{d}{dx} \left\{ [G(x - \xi) - G(t - \xi)] \frac{d}{d\xi} G^{*j}(\xi) \right\} d\xi dt$$

and, assuming that $\frac{d}{dx} G(x)$ equals $g(x)$,

$$\frac{d}{dx} F(x) = \sum_{j=0}^{\infty} \int_0^x \int_0^t \mu e^{-\mu t} g(x - \xi) dG^{*j}(\xi) dt$$

Next, by interchanging the order of integration and revising the limits appropriately, we get

$$\begin{aligned} \frac{d}{dx} F(x) &= \sum_{j=0}^{\infty} \int_0^x \left[\int_{\xi}^x \mu e^{-\mu t} dt \right] g(x - \xi) dG^{*j}(\xi) \\ &= \sum_{j=0}^{\infty} \int_0^x \left[e^{-\mu \xi} - e^{-\mu x} \right] g(x - \xi) dG^{*j}(\xi) \end{aligned}$$

Therefore, the Laplace-Stieltjes transform, $\gamma(z)$, is as follows

$$\begin{aligned}
\gamma(z) &= \int_0^{\infty} e^{-zx} dF(x), \quad \operatorname{Re} z \geq 0. \\
&= \sum_{j=0}^{\infty} \int_0^{\infty} \int_0^x e^{-zx} \left[e^{-\mu\xi} - e^{-\mu x} \right] g(x-\xi) dG^{*j}(\xi) dx
\end{aligned}$$

Interchanging the order of integration once again, we get

$$\gamma(z) = \sum_{j=0}^{\infty} \int_0^{\infty} \int_{\xi}^{\infty} \left\{ e^{-zx} \left[e^{-\mu\xi} - e^{-\mu x} \right] g(x-\xi) \right\} dx dG^{*j}(\xi)$$

Now, the following change of variables is made,

$$v = x - \xi, \quad dx = dv,$$

such that

$$\begin{aligned}
\gamma(z) &= \sum_{j=0}^{\infty} \int_0^{\infty} \int_0^{\infty} e^{-z(v+\xi)} \left[e^{-\mu\xi} - e^{-\mu(v+\xi)} \right] g(v) dv dG^{*j}(\xi), \\
&= \sum_{j=0}^{\infty} \int_0^{\infty} e^{-(z+\mu)\xi} \int_0^{\infty} \left[e^{-zv} - e^{-(z+\mu)v} \right] g(v) dv dG^{*j}(\xi).
\end{aligned}$$

Remembering that

$$\begin{aligned}
\Psi(z) &= \int_0^{\infty} e^{-zv} dG(v) \\
&= \int_0^{\infty} e^{-zv} g(v) dv,
\end{aligned}$$

we have

$$\gamma(z) = \sum_{j=0}^{\infty} \int_0^{\infty} e^{-(z+\mu)\xi} [\Psi(z) - \Psi(z+\mu)] dG^{*j}(\xi)$$

Finally, we take $[\Psi(z) - \Psi(z+\mu)]$ outside of the integral and summation signs to get

$$\gamma(z) = [\Psi(z) - \Psi(z+\mu)] \sum_{j=0}^{\infty} \int_0^{\infty} e^{-(z+\mu)\xi} dG^{*j}(\xi),$$

and, since the transform of the j -fold convolution of a distribution function equals the transform of that distribution function raised to the j^{th} power, we have

$$\begin{aligned} \gamma(z) &= [\Psi(z) - \Psi(z+\mu)] \sum_{j=0}^{\infty} \Psi^j(z+\mu), \\ &= \frac{\Psi(z) - \Psi(z+\mu)}{1 - \Psi(z+\mu)}, \quad \operatorname{Re} z \geq 0 \end{aligned}$$

Since it is true that

$$E(X) = - \frac{d}{dz} \gamma(z) \Big|_{z=0},$$

we must differentiate $\gamma(z)$ with respect to z . First,

$$\begin{aligned} \frac{d}{dz} \Psi(z+\mu) &= \frac{d}{d(z+\mu)} \Psi(z+\mu), \\ &= \Psi'(z+\mu). \end{aligned}$$

Therefore

$$-\frac{d}{dz} \gamma(z) = -\frac{[1 - \Psi(z+\mu)][\Psi'(z) - \Psi'(z+\mu)] + \Psi'(z+\mu)[\Psi(z) - \Psi(z+\mu)]}{[1 - \Psi(z+\mu)]^2}$$

Setting $z = 0$ and noting that $\Psi(0) = 1$,

$$\begin{aligned} -\frac{d}{dz} \gamma(z) \Big|_{z=0} &= E(X) , \\ &= -\frac{[1 - \Psi(\mu)][\Psi'(0) - \Psi'(\mu)] + \Psi'(\mu)[1 - \Psi(\mu)]}{[1 - \Psi(\mu)]^2} \\ &= -\frac{\Psi'(0) - \Psi'(\mu) + \Psi'(\mu)}{1 - \Psi(\mu)} \\ &= \frac{-\Psi'(0)}{1 - \Psi(\mu)} . \end{aligned}$$

Now, adding back the subscript, we have shown that the expected cycle length for the m^{th} server is

$$E(X_m) = \frac{-\Psi'_m(0)}{1 - \Psi_m(\mu_m)}$$

Therefore

$$p_m = -\frac{\tau_m [1 - \Psi_m(\mu_m)]}{\Psi'_m(0)} ,$$

and from equation 5 of Chapter IV,

$$\frac{-1}{\Psi'_m(0)} = \begin{cases} \lambda & m = 1 \\ \lambda \prod_{i=1}^{m-1} \Psi_i(\mu_i) & m = 2, 3, \dots, s \end{cases}$$

so that

$$p_m = \begin{cases} \frac{\lambda}{\mu_m} [1 - \Psi_m(\mu_m)] & m = 1 \\ \frac{\lambda}{\mu_m} \left[\prod_{i=1}^{m-1} \Psi_i(\mu_i) \right] [1 - \Psi_m(\mu_m)] & m = 2, 3, \dots, s. \end{cases}$$

This is equation 4.9, and so the assertion is proved.

REFERENCES

1. C. J. Ancker, Jr. and A. V. Gafarian, "Queueing with Reneging and Multiple Heterogeneous Servers," Naval Research Logistics Quarterly, 10, 125-149 (1963).
2. Tom M. Apostol, Mathematical Analysis, Addison Wesley Publishing Co., Inc., Reading, Mass., 1957
3. Vaclav E. Benes, "On Trunks with Negative Exponential Holding Times Serving a Renewal Process," Bell System Technical Journal, 38, 211-258 (Jan., 1959).
4. Paul J. Burke, "The Overflow Distribution for Constant Holding Time," Bell System Technical Journal, 50, 3195-3210, (Dec., 1971).
5. Erhan Cinlar and Ralph Disney, "Stream of Overflows from a finite Queue," Operations Research, 15, 131-134, (1967).
6. Robert B. Cooper, Introduction to Queueing Theory, Macmillan Co., New York, N.Y., 1972.
7. Robert B. Cooper, "Queues with Ordered Servers that Work at Different Rates." To be presented at the 20th International Meeting of the Institute of Management Sciences, Tel Aviv, Israel, June 24-29, 1973.
8. A. Descloux, "On Markovian Servers with Recurrent Input," Proceedings of the Sixth International Teletraffic Congress, Munich, Sept. 9-15, 1970.
9. A. Descloux, "On Overflow Processes of Trunk Groups with Poisson Input and Exponential Service Times," Bell System Technical Journal, 42, 383-398 (March, 1963).
10. Harold Gumbel, "Waiting Lines with Heterogeneous Servers," Operations Research, 8, 504-511 (July-Aug. 1960).
11. P. G. Hoel, S. C. Port, and C. J. Stone, Introduction to Stochastic Processes, Houghton Mifflin Co., Boston, 1972.
12. A. Y. Khintchine, Mathematical Methods in the Theory of Queueing, 2nd ed. Hafner Publishing Co., New York, N.Y., 1969.

13. Michael G. Kolfrath and Arthur L. Smith, "On a Multiple Exponential Channel Service Facility with Heterogeneous Mean Service Rates," Scientific and Technical Aerospace Reports, Accessions No. N69-12516 (1969).
14. B. Krishnamoorthi, "On Poisson Queue with Two Heterogeneous Servers," Operations Research, 11, 321-330 (May-June, 1963).
15. C. Palm, "Intensitatsschwankungen im Fernspreverkehr," Ericsson Technics, 44, 1-189 (1943).
16. John Riordan, Stochastic Service Systems, John Wiley & Sons, Inc., New York, N.Y., 1962.
17. Vijendra P. Singh, "Two Server Markovian Queues with Balking: Heterogeneous vs Homogeneous Servers," Operations Research, 18, 145-159 (Jan.-Feb., 1970).
18. Vijendra P. Singh, "Markovian Queues with Three Heterogeneous Servers," A.I.I.E. Transactions, 3, 45-48 (March, 1971).
19. F. Spitzer, "Random Fields and Interacting Particle Systems," Notes on lectures given at a summer seminar of the Mathematical Association of America at Williams College, Williamstown, Mass., 1971.
20. R. Syski, Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd, Ltd., London, 1960.
21. Lajos Takács, Introduction to the Theory of Queues, Oxford University Press, New York, N.Y., 1962.
22. D. G. Tambouratzis, "Limiting Distribution of the Intervals between Losses when the Input is Recurrent and the Service Times Negative Exponential," Scientific and Technical Aerospace Reports, Accessions No. N70-13723 (1970).
23. E. Vaultot, "Application du calcul des probabilités à l'exploitation téléphonique," Annales des Postes, Télégraphes et Téléphones Vol. 14, No. 2 (1925), pp. 136-156.